Weekly Report

Junhua Lu

May 17, 2015

About the data The data contains records of 14 million permanent resident(常住人口), 10 million temporary resident(暂住人口), and several thousands of monitored people. The *monitored people* means people who had committed crimes but have been released(but they are still harmful to society). The former two kinds of people includes criminals and normal people.

The following are the attributes. However many attributes are not included, because Gu TY thought they are not useful for our study. Like hotel, internet cafe, marital status and occupation, they are stored in the database(in one software PLSQL on the disk in Qiqiaoban). So these data should be retrieved from the database and then store them in CSV files later.

数据表字段

2015年4月22日 星期三 15:06

100w张表的字段是
编号. 身份证GMSFHM. 性别XB. 职业ZY. 文化程度WHCD. 出生地省市县区CSDSSXQ

Zzf和10w
编号. 身份证GMSFHM. 性别XB. 职业ZY. 文化程度WHCD. 出生地省市县区CSDSSXQ. 户号HH

Hotel
证件号码ZJHM

Pre1920以及其他年份表
公民身份号码GMSFHM. 出生日期CSRQ. 监护人一公民身份号码JHRYGMSFHM.
监护人二公民身份号码JHREGMSFHM. WHCD文化程度.
与户主关系YHZGX. 户号HH. 街路巷JLX.

Algorithms and models Several methods may be applicable in this project.

- Linear Regression(since the output is 0,1 Logistic Regression is better), Generalized Linear Mixed Modal. Put the data in, and regression.
- Decision Tree.
- Classifiers like Support Vector Machine and Nearest Neighbouring.
- Bayesian network, dynamic Bayesian network, TV-DBN. Hidden Markov Model maybe included with DBN. In Zhu MF's opinion, we may cluster the people recorded, find the cluster with high rate of commit crimes HR and compute the transition matrices to see which cluster has high possibility to *transform* into HR. It may be a good method, and I am trying to get more familiar with the DBN methods and discuss about this tomorrow.
- Conditional Random Field. I'm not familiar with this one. Will learn it soon.
- Duration Model. As mentioned before, the R function for estimating the parameters does not convergent, but it is a good model for this kind of data.

Next week

1. Speech at group meeting on Thursday.

2. Learn more about above methods and implement them(linear regression,and dbn or SVM) after data processing.